Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning

Year: 2023 | *Citations: 313* | *Authors: Antoine Yang, Arsha Nagrani, P. H. Seo*

Abstract

In this work, we introduce Vid2Seq, a multi-modal single-stage dense event captioning model pretrained on narrated videos which are readily-available at scale. The Vid2Seq architecture augments a language model with special time tokens, allowing it to seamlessly predict event boundaries and textual descriptions in the same output sequence. Such a unified model requires large-scale training data, which is not available in current annotated datasets. We show that it is possible to leverage unlabeled narrated videos for dense video captioning, by reformulating sentence boundaries of transcribed speech as pseudo event boundaries, and using the transcribed speech sentences as pseudo event captions. The resulting Vid2Seq model pretrained on the YT-Temporal-1B dataset improves the state of the art on a variety of dense video captioning benchmarks including YouCook2, ViTT and ActivityNet Captions. Vid2Seq also generalizes well to the tasks of video paragraph captioning and video clip captioning, and to few-shot settings. Our code is publicly available at [1].